REANA and ML

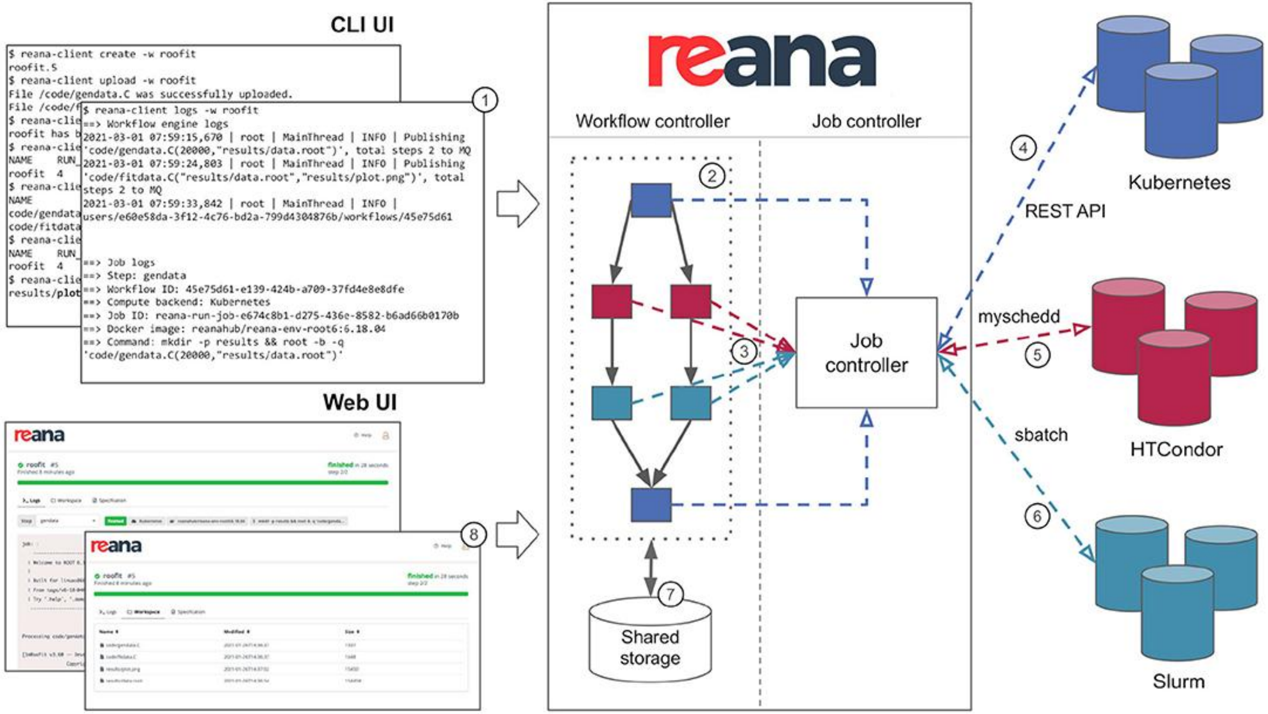
Tibor Simko
IT-PW

CERN IT Machine Learning Infrastructure Workshop, March 10th 2023

https://indico.cern.ch/event/1253881

REANA Reusable Analysis platform

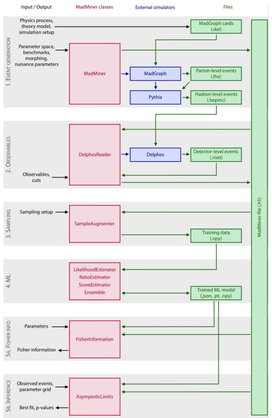
https://www.reana.io

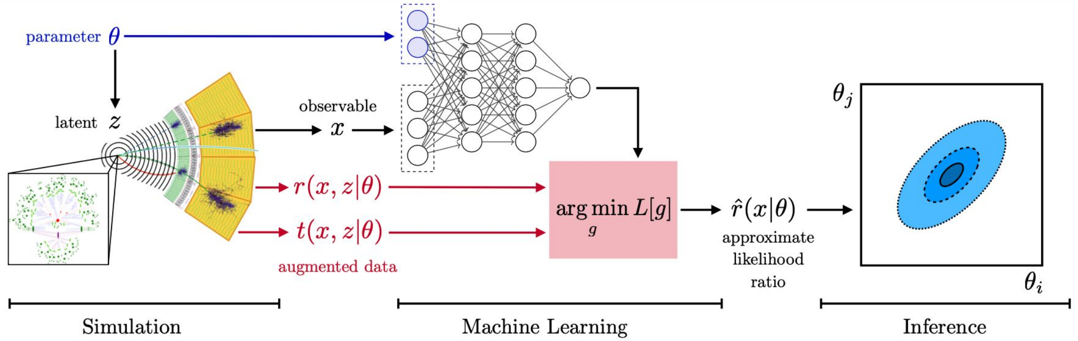


Running declarative containerised computational workflows

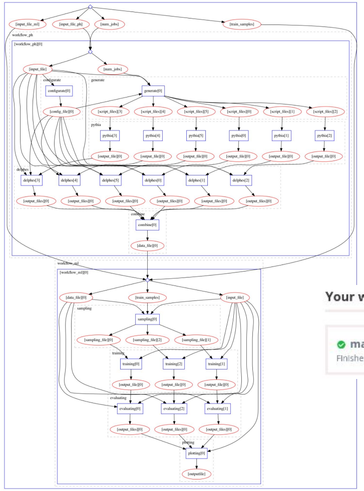
ML use cases on REANA 1/2

Pheno-level analyses embedded into Python ML ecosystem (and optionally MLFlow)

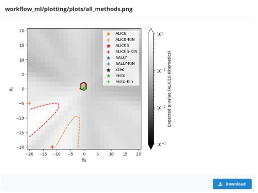




"MadMiner: Machine learning-based inference for particle physics", J. Brehmer, F. Kling, I. Espejo, K. Cranmer, arXiv:1907.10621.

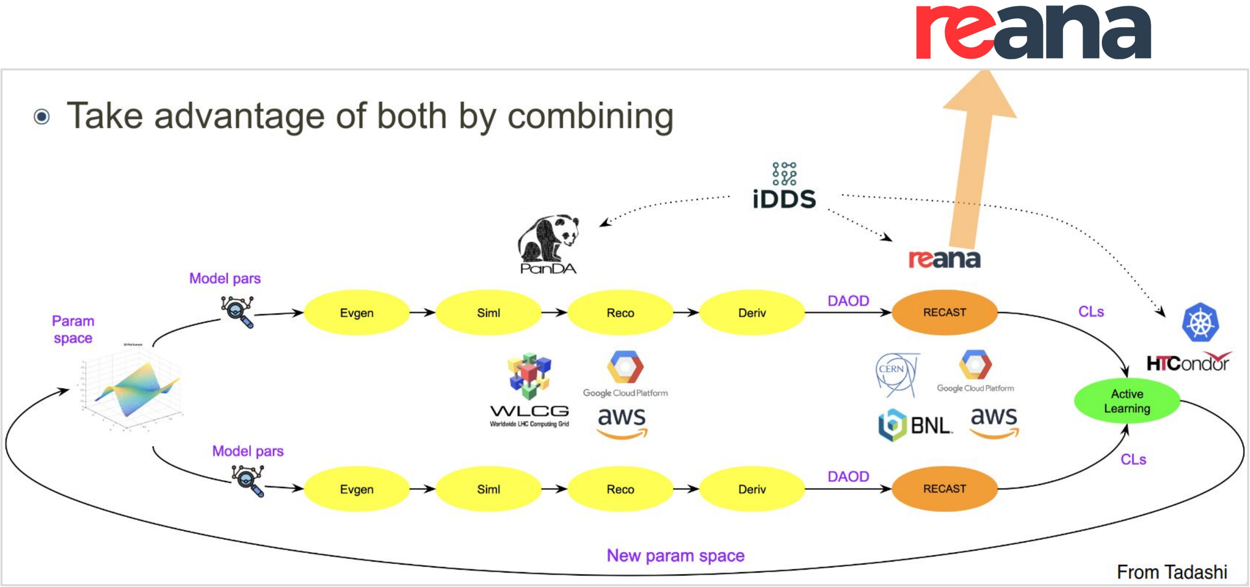


reana

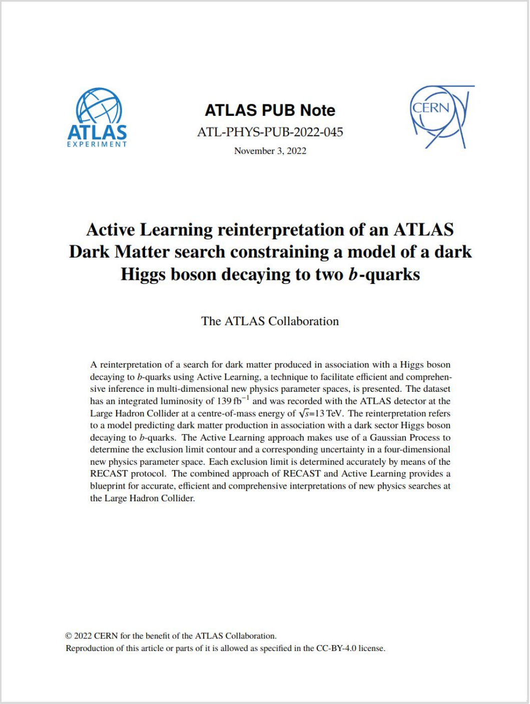


- Running ML based workflows

ML use cases on REANA 2/2



“REANA / PanDA integration for Active Learning”, W.Guan, T.Maeno, C.Weber, T.Wenaus, R.Zhang, https://indico.cern.ch/event/1134581.



ATLAS PUB Note
ATL-PHYS-PUB-2022-045
November 3, 2022

Active Learning reinterpretation of an ATLAS Dark Matter search constraining a model of a dark Higgs boson decaying to two b-quarks

The ATLAS Collaboration

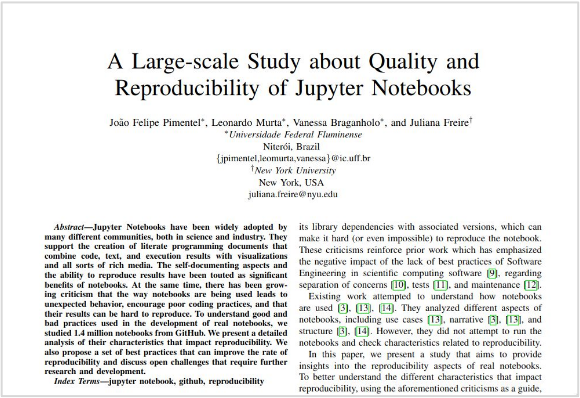
A reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to b-quarks using Active Learning, a technique to facilitate efficient and comprehensive inference in multi-dimensional new physics parameter spaces, is presented. The dataset has an integrated luminosity of 139 fb^{-1} and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s}$=13 TeV. The reinterpretation refers to a model predicting dark matter production in association with a dark sector Higgs boson decaying to b-quarks. The Active Learning approach makes use of a Gaussian Process to determine the exclusion limit contour and a corresponding uncertainty in a four-dimensional new physics parameter space. Each exclusion limit is determined accurately by means of the RECAST protocol. The combined approach of RECAST and Active Learning provides a blueprint for accurate, efficient and comprehensive interpretations of new physics searches at the Large Hadron Collider.

ATL-PHYS-PUB-2022-045

- Running workflows as part of a bigger data processing chain (whole physics analysis from MC generation to new physics discovery)

Possible areas of interest

- Capturing the knowledge behind data analyses
  → *preserve to reuse*
- Computational reproducibility
  → *run outside the original context*
- Running workflows at scale
  → *10k workflows for ATLAS pMSSM searches*
- “Continuous analyses”
  → *Gitlab-REANA bridge*
- Interplay between notebooks and workflows
  → *interactive vs batch*



A Large-scale Study about Quality and Reproducibility of Jupyter Notebooks

João Felipe Pimentel*, Leonardo Murta*, Vanessa Braganholo*, and Juliana Freire†
*Universidade Federal Fluminense
Niterói, Brazil
{jpimentel,leomurta,vanessa}@ic.uff.br
†New York University
New York, USA
juliana.freire@nyu.edu

***Abstract*—Jupyter Notebooks have been widely adopted by many different communities, both in science and industry. They support the creation of literate programming documents that combine code, text, and execution results with visualizations and all sorts of rich media. The self-documenting aspects and the ability to reproduce results have been touted as significant benefits of notebooks. At the same time, there has been growing criticism that the way notebooks are being used leads to unexpected behavior, encourage poor coding practices, and that their results can be hard to reproduce. To understand good and bad practices used in the development of real notebooks, we studied 1.4 million notebooks from GitHub. We present a detailed analysis of their characteristics that impact reproducibility. We also propose a set of best practices that can improve the rate of reproducibility and discuss open challenges that require further research and development.**

***Index Terms*—jupyter notebook, github, reproducibility**

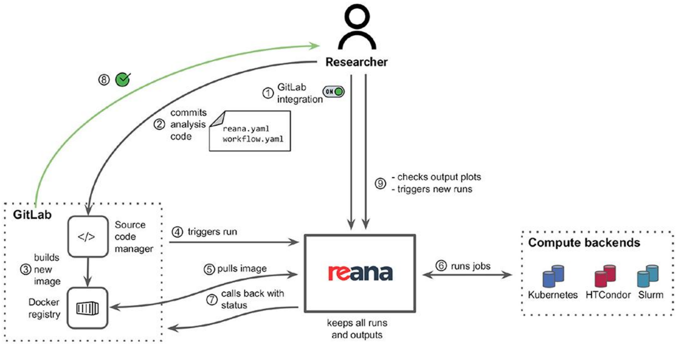
its library dependencies with associated versions, which can make it hard (or even impossible) to reproduce the notebook. These criticisms reinforce prior work which has emphasized the negative impact of the lack of best practices of Software Engineering in scientific computing software [9], regarding separation of concerns [10], tests [11], and maintenance [12].

Existing work attempted to understand how notebooks are used [3], [13], [14]. They analyzed different aspects of notebooks, including use cases [13], narrative [3], [13], and structure [3], [14]. However, they did not attempt to run the notebooks and check characteristics related to reproducibility.

In this paper, we present a study that aims to provide insights into the reproducibility aspects of real notebooks. To better understand the different characteristics that impact reproducibility, using the aforementioned criticisms as a guide,

“...only 4.03% produced the same results”

DOI 10.1109/MSR.2019.00077



GitLab-REANA bridge